

PROBLEM STATEMENT

 Suppose you love to read fantasy books, and you want to be updated on book reviews and the latest book releases...

How do we build a model capable of understanding our likes, and then filtering incoming posts to show us what we want?

METHODOLOGY

Datasets

- 1800 posts from science fiction and fantasy subreddits from Jan 27 28
- Cross reference datasets with reputable sources
- Fix errors

Data exploration phase & pre-processing

- Remove duplicate posts
- Remove posts with insufficient content
- Removing html, punctuation, stop-words

Modelling

- Naïve Bayes
- Logistic Regression

KEY FINDINGS

FROM DATA

Naive Bayes model scored

86%

Logistic Regression model scored

85%

```
new evenfirst best author hope thank never say finish take person thought live may be to magic manifer lot work start happen novel to make thought live may be to make thought live may be to make thought live may be to make the power thing was a come idea help research to the make part human recommendance want somethe still book show
```

Top words from training dataset

CHALLENGES

■ Duplicate posts → switched reddits

■ Overfitting → Customised stopwords